DOCUMENT RESUME

ED 242 788 TM 840 207

AUTHOR	Thissen, David; Steinberg, Lynne
TITLE	A Response Model for Multiple Choice Items. Psychometric Technical Report No. 1.
INSTITUTION	National Opinion Research Center, Chicago, Ill.
SPONS AGENCY	Office of Naval Research, Arlington, Va. Personnel and Training Research Programs Office.
PUB DATE	Dec 83
CONTRACT	N00014-83-C-0283
NOTE	52p.
PUB TYPE	Reports - Research/Technical (143)
EDRS PRICE	MF01/PC03 Plus Postage.
DESCRIPTORS	Ability; Estimation (Mathematics); *Guessing (Tests); *Latent Trait Theory; *Mathematical Models; *Maximum Likelihood Statistics; *Multiple Choice Tests; *Responses; Testing; Test Items
IDENTIFIERS	Armed Services Vocational Aptitude Battery; *Item Parameters

ABSTRACT

An extension of the Bock-Samejima model for multiple choice items is introduced. The model provides for varying probabilities of the response alternative when the examinee guesses. A marginal maximum likelihood method is devised for estimating the item parameters, and likelihood ratio tests for comparing more and less constrained forms of the model are provided. Applications of the model are illustrated with item response data for the word knowledge and general science subtests of the Armed Services Vocational Aptitude Battery. (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED242788

A Response Model for Multiple Choice Items

David Thissen
and
Lynne Steinberg
University of Kansas

Psychometric Technical
Report No. 1

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

The Office of Naval Research

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

A Response Model for
Multiple Choice Items

David Thissen
and
Lynne Steinberg
University of Kansas

Psychometric Technical
Report No. 1

This research was sponsored by the Personnel and Training Research Program, Psychological Sciences Division, Office of Naval Research under Contract No. N00014-83-C-0283, Contract Authority Identification No. NR 475-018.

Approved for public release; distribution unlimited. Reproduction in whole or part is permitted for any purpose of the United States Government.

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER Psychometric Technical Report No. 1	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) A Response Model for Multiple Choice Models		5. TYPE OF REPORT & PERIOD COVERED 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) David Thissen and Lynne Steinberg		8. CONTRACT OR GRANT NUMBER(s) N00014-83-C-0283
9. PERFORMING ORGANIZATION NAME AND ADDRESS NORC 6030 S. Ellis Avenue Chicago, IL 60637		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 62763N RF63521803 RF63521 NR 475-018
11. CONTROLLING OFFICE NAME AND ADDRESS Personnel and Training Research Programs Office of Naval Research (Code 442PT) Arlington, VA 22217		12. REPORT DATE December 1983 13. NUMBER OF PAGES 43
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

An extension of the Bock-Samejima model for multiple choice items is introduced. The model provides for varying probabilities of the response alternative when the examinee guesses. A marginal maximum likelihood method is devised for estimating the item parameters, and likelihood ratio tests for comparing more and less constrained forms of the model are provided. Applications of the model are illustrated with item response data for the word knowledge and general science subtests of the Armed Services Vocational Aptitude Battery.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-LF-014-6601

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

A RESPONSE MODEL FOR MULTIPLE CHOICE ITEMS

David Thissen and Lynne Steinberg

University of Kansas

ABSTRACT

An extension of the Bock-Samejima model for multiple choice items is introduced. The model provides for varying probabilities of the response alternative when the examinee guesses. A marginal maximum likelihood method is devised for estimating the item parameters, and likelihood ratio tests for comparing more and less constrained forms of the model are provided. Applications of the model are illustrated with item response data for the word knowledge and general science subtests of the Armed Services Vocational Aptitude Battery.

A Response Model for Multiple Choice Items

In practical ability measurement, binary item response models have been applied routinely to data from multiple-choice tests with four or five alternatives per item. To use the binary models, the data have been dichotomized (correct and incorrect) and the distinct identity of the incorrect alternatives has been lost. This procedure logically follows the tradition of "scoring" a test by considering only the correct answers; however, the binary models are incomplete as theories of the item responses as useful information may be lost in the dichotomization.

It is possible to conceive of a unidimensional latent variable model which completely explains the item response data, and several have been proposed. The model which will be introduced here is an extension of that proposed by Bock (1972) and subsequently extended by Samejima (1979); a different parameterization of the same model has been discussed by Sympson (1983). Bock (1972), Thissen (1976), Sympson (1983), and Levine and Drasgow (1983) have shown that some increased precision of measurement may be obtained when information in the incorrect alternatives on multiple choice tests is included in the item response model. In a subsequent section, we will discuss this gain in information with a new model, as well as some attendant difficulties. In the next section we will introduce the model and describe a workable scheme for maximum likelihood estimation of its parameters.

The Model

The model for multiple-choice items to be described makes heavy use of the multivariate logistic transformation suggested as an item response model by Bock (1972), so it is useful to begin with a clear understanding of that model. For m categorical responses, we specify m response functions $z_k = a_k\theta + c_k$, each of which is a linear function of the latent ability variable θ. Usually $z_{correct}$ would have a positive slope and the other z_k's would have lesser positive or negative slopes. The linear functions describe one of the simplest possible relationships between the response and θ. To make this a model for categorical item responses, the z_k (which lie on the real line) must be mapped into [0,1]. This is accomplished by the multivariate logistic transformation, so

$$P(x_j=h|\theta;\underline{a},\underline{c}) = \frac{\exp(z_h)}{\sum_{k=1,m_j} \exp(z_k)} \quad . \quad (1)$$

The function (1) is Bock's (1972) model for an item response, $x_j=h$, in which $h = 1,2,\ldots m_j$ for a multiple choice item j with m_j (classes of) response alternatives. The item parameters are the vectors $\underline{a}$ and $\underline{c}$, subject to two suitable linear constraints (see below), giving $2(m_j-1)$ free parameters. The model described by (1) is moderately flexible, but lacks flexibility in certain crucial respects at the extremes. Specifically, one of the response alternatives must have the largest positive value of a_k; the trace line for that alternative is then monotonic increasing. That may be theoretically acceptable, since it

is probably the correct response. However, another alternative must have the largest negative value of a_k ($\Sigma a_k=0$ is one of the linear constraints), and that alternative's trace line must be monotonic decreasing. The latter aspect of the model is less acceptable: it implies that as ability decreases the probability of selecting one particular incorrect response approaches unity, and all the others go to zero. That is unlikely.

Samejima (1979) proposed a solution to this problem in a conceptual modification of Bock's (1972) model in which she added a completely latent response category labelled "zero." Here, we will sometimes refer to this category as "don't know" (DK). Lord (1982) has introduced a similar conceptual entity which he describes by saying the examinee is "(totally) undecided." In Samejima's model, DK is not an observed response, but a latent one, multi-logit-linearly dependent on a latent trait (giving two layers latent); the idea was that some proportion d_h ($=1/m_j$ in Samejima (1979)) "guessed" each of the observable response alternatives and were "mixed in" with the examinees who chose those alternatives intentionally. So the model becomes

$$P(x_j=h|\theta;\underline{a},\underline{c},\underline{d}) = \frac{\exp(z_h)}{\sum_{k=0,m_j}\exp(z_k)} + \frac{d_h\ \exp(z_0)}{\sum_{k=0,m_j}\exp(z_k)} ,$$

in which the second term adds a proportion d_h of those who "don't know" into each trace line; call it $P_j(h)$ for future brevity:

$$P_j(h) = \frac{\exp(z_h) + d_h\ \exp(z_0)}{\sum_{k=0,m_j}\exp(z_k)} , \qquad (2)$$

in which h takes the values $1,2,\ldots m_j$. Samejima's (1979) model had two more parameters than Bock's (1972) model: a_0 and c_0. In her presentation, the d_h were fixed and equalled $1/m_j$; this represents the hypothesis that those of sufficiently low ability assign their responses randomly with equal probability to each of the response alternatives. We found that unlikely; later we will show that it is not empirically the case. So we extend the version of the model given in (2) to allow the d_h, $h=1,2,\ldots m_j$ to be functions of estimated parameters.

Indeterminacies and Constraints

The model expressed in (2) has a number of indeterminacies and requires the imposition of some constraints to become identifiable. The linear constraints required on $\underline{a}$ and $\underline{c}$,

$$\sum_{k=0,m_j^k} a_k = \sum_{k=0,m_j^k} c_k = 0$$

are imposed by reparameterization:

$$\underline{a} = T\underline{\alpha}$$

and

$$\underline{c} = T\underline{\gamma} \quad ,$$

where T is an $(m_j+1) \times m_j$ transformation matrix, and $\underline{\alpha}$ and $\underline{\gamma}$ are vectors of free parameters of order m_j; T is the transpose of that in Bock (1972) throughout. There is a further indeterminacy with respect to the sign of $\underline{\alpha}$; this is an indeterminacy of reflection of the latent variable

θ which is logically identical to the rotational indeterminacy of common factor analysis. The reflection indeterminacy is not explicity solved, but rather the iterative solution of the likelihood equations is started with a_h positive for h = the correct response; this is usually sufficient to keep the estimated solution "right-side-up."

Since the vector $\underline{d}$ represents a set of m_j proportions, the d_h must be constrained to sum to unity and lie on the interval [0,1]. The dual constraint is imposed in two parts. In the first, the same multivariate logistic used in the model is also employed to make the d_h proportions from a set of psuedo-parameters d_h^*, which lie on the real line:

$$d_h = \exp(d_h^*) \; / \sum_{k=1,m_j} \exp(d_k^*) \quad . \qquad (3)$$

The parameters d_h^* are on the real line, but like $\underline{a}$ and $\underline{c}$, they must satisfy $\Sigma d_k^* = 0$. So

$$\underline{d}^* = T_2\underline{\delta} \qquad (4)$$

in which T_2 is an m_j x (m_j-1) transformation matrix of the same form as T (above) but of lesser order, and $\underline{\delta}$ is a vector of length (m_j-1) of parameters which (through (4) and (3)) give the proportions d_h.

So the set of free parameters for an item consists of m_j α_k's ("a-contrasts"), m_j γ_k's ("c-contrasts"), and (m_j-1) δ_k's ("d-contrasts"), for a total of $3m_j-1$ parameters: 11 for a four-alternative multiple-choice item.

The estimation system we use permits any of the elements of $\{\underline{\alpha},\underline{\gamma},\underline{\delta}\}$ to be fixed at any constant value, or to be constrained to be

equal to any other parameter. With this facility, our model (2) includes the previous models hierarchically as subsets. If $\underline{\delta}$ is fixed at $\underline{0}$, than all the $d_h=1/m_j$ and (2) becomes Samejima's (1979) model. If the parameters a_0 and c_0 are fixed at zero, the d_h become irrelevant (and are deleted mechanically) and the model becomes Bock's (1972) model. In our examples, we will compare the goodness of fit of all of these models, as well as versions of (2) that include equality constraints on particular parameters imposed across items.

The Relationship of the Model to Data

Models in item response theory (IRT) are intended to explain observed covariation among test-item responses. Unidimensional "latent trait" theories explain that covariation by appeal to an underlying latent variable (usually denoted θ) on which the probabilities of the responses are functionally dependent. Each pair of item responses is theorized to have non-zero covariance because the probability of both item responses depends on θ. The models usually hold that θ is the sole cause of covariation among the item responses; conditional on θ, the item responses are theorized to be independent ("local independence").

A complete IRT model for the item response data for a multiple-choice test in which n items each have m alternatives would be a model for the covariation in the m^n contingency table containing the counts of respondents giving each response pattern. Such a 4^2 table for two

four-alternative multiple-choice items is given in Table 1.

Insert Table 1 About Here

Evidence that there is some information in the examinees' incorrect choices is apparent (or at least can be found) in the covariance in the table. (The column percentages for the counts are given in parentheses to aid in interpretation.) Alternative C is correct for item 2; respondents who select C on item 2 are most likely to choose B on item 1; B is correct, so that is double evidence of their high ability. But examinees who choose A or B on item 2 are more likely to respond correctly to item 1 (36% and 34%) than those who pick D on item 2 (only 25%). A possible explanation is that those who select A or B on item 2 are higher on the ability continuum than are those who pick D. Even with 16 cells, this sort of argument may be lengthy, but the idea is that wrong responses are related to each other; if the responses are locally independent, this implies a relationship to "the trait."

The goodness of fit of a multiple category IRT model to the data in such a table may be evaluated with conventional likelihood ratio χ^2 statistics. For sets of data with more examinees than cells in the table, the likelihood ratio test against a general multinomial alternative may be used. For larger tables (35 four-choice items gives a 4^{35} table), there is no "general multinomial alternative"; but the likelihood ratio test between hierarchically nested models still may be used to evaluate the significance of the additional parameters of the

larger model.

Estimation

In general, the data for the estimation of the parameters $\{\underline{\alpha},\underline{\gamma},\underline{\delta}\}_j$ for each item j, j=1,...,n consist of the counts of response patterns $r_{\underline{x}}$ (where each x_j may take values in $[1,2,\ldots m_j]$) in an $m_1 \times m_2 \times \ldots \times m_n$ contingency table like Table 1. The probability of observing a particular response pattern $\underline{x}$ when drawing an examinee from a population in which θ follows a distribution $\phi(\theta)$, assumed N(0,1) in the examples, is

$$P(\underline{x}) = \int_{-\infty}^{\infty} \prod_{j=1,n} P_j(x_j)\ \phi(\theta)\ d\theta \qquad (5)$$

in which $P_j(x_j)$ is from equation (2).

The likelihood for the entire set of observed data is

$$L = C \prod_{\underline{x}} P(\underline{x})^{r_{\underline{x}}} \qquad (6)$$

in which C is a normalizing constant not dependent on the parameters, and the product runs over all possible response patterns. In practice, of course, for more than a handful of items the number of possible response patterns is astronomical, and the "count" in each cell is one or zero. In such cases, only those response patterns actually observed need to be considered in the computation of (6), or (7), a function proportional to the loglikelihood:

$$\ell_o \sim \sum_{\underline{x}} r_{\underline{x}} \log P(\underline{x}) \quad . \qquad (7)$$

For a very small number of items (3 or 4; the parameters of the full model are not identifiable with only 2 items, as there are fewer cells than parameters), it may be possible to obtain the MLEs by directly maximizing (7) with a Newton-type algorithm. Bock (1972) reported such results for his original model with procedures similar to those used by Bock and Lieberman (1970).

Bock and Aitkin (1981) described an extremely simple and elegant algorithm for maximizing functions like (7) for binary models, and we extend that algorithm to the estimation of multiple category models. The Bock-Aitkin algorithm is a two-step procedure like the "EM-algorithm" (Dempster, Laird, and Rubin, 1977), so we describe first the "E-step" and then the "M-step." The procedure is iterative, and EM-pairs are repeated until the process converges.

The E-step

The Bock-Aitkin algorithm is based on a discrete representation of $\phi(\theta)$ and the integrand of $P(\underline{x})$, both continuous densities, over Q "quadrature points" θ_q, with q=1,2,...Q. Such a discrete representation of the continuous densities may be made arbitrarily close to continuous reality by choosing Q large, just as numerical integration may be made arbitrarily accurate by using sufficient quadrature points. However, large values of Q slow the computations; Q=10 over the range θ=-4.5 to θ=4.5 in unit steps seems to be sufficiently accurate and fast for many applications and is used in all of the examples here.

Under the assumption that the population is composed of individuals who are members of Q discrete "classes" with values θ_1, θ_2, ..., θ_Q on the latent variable, "complete data sufficient statistics" for the estimation of the item parameters $\{\underline{\alpha},\underline{\gamma},\underline{\delta}\}_j$ for item j would consist of a table of counts R_j^*, in which each element r^*_{jkq} is the number of individuals in class θ_q selecting response alternative k on item j. So the E-step of the Bock-Aitkin algorithm consists of computing the expected values of the r^*_{jkq}, conditioned on the data and the current provisional estimates of the item parameters.

Using provisional estimates of the item parameters for each item (as starting values we use $\underline{\delta}=\underline{0}$, $\underline{\gamma}=\underline{0}$, and $\alpha_k=1$ for categories k which are incorrect and $\alpha_k=2$ for category k correct), compute the elements of the m_j x Q table R_j^* containing:

$$E(r^*_{jkq} \mid \text{data}; \{\hat{\underline{\alpha}},\hat{\underline{\gamma}},\hat{\underline{\delta}}\}) = \sum_{(\underline{x} \text{ in } \xi)} r_{\underline{x}} [P(\underline{x};\theta_q)/\Sigma_q P(\underline{x};\theta_q)] \quad (8)$$

in which ξ is the set of $\underline{x}$ in which $x_j=k$ and

$$P(\underline{x};\theta_q) = \prod_{j=1,n} P_j(x_j)\phi(\theta_q) \quad . \quad (9)$$

Note that, while (9) and (8) are computed in a (potentially long) loop over the observed response patterns, the values $P_j(k)$ for each item from (2) are required only for a fixed set of Q values of θ. If those values are placed in a table before the E-step is begun, the computations involved in (8) and (9) are limited to table look-up, multiplication, and addition. The E-step yields a set of n m_j x q tables of non-integral artificial "counts" which are used as data in the M-step.

The M-step

The M-step consists of maximum likelihood estimation of the parameters $\{\underline{\alpha},\underline{\gamma},\underline{\delta}\}_j$ for all items j=1,2,...n, using the tables of expected values R_j^* as data. It is simply nonlinear regression.

In terms of the "data" in R_j^*, the loglikelihood for item j is

$$\ell_j \sim \sum_k \sum_q r^*_{jkq} \log P_j(k;\theta_q) \qquad (10)$$

in which $P_j(k;\theta_q)$ is equation (2) evaluated at θ_q. Standard gradient methods may be used to maximize ℓ_j over the parameter space. The MLEs of the parameters are obtained where

$$\frac{\partial \ell_j}{\partial \zeta} = 0$$

for all parameters in $\zeta=\{\underline{\alpha},\underline{\gamma},\underline{\delta}\}$. Since for any set of parameters constrained by T such that $\kappa=T\zeta$,

$$\frac{\partial \ell_j}{\partial \zeta} = T \frac{\partial \ell_j}{\partial \kappa} \quad , \qquad (11)$$

we will state the required components of the gradient for the constrained parameters $\{\underline{a},\underline{c},\underline{d}\}$; the gradients used to maximize (10) are then given by (11). Further,

$$\frac{\partial \ell_j}{\partial \kappa} = \sum_k \sum_q r^*_{jkq} \frac{1}{P_j(k;\theta_q)} \frac{\partial P_j(k;\theta_q)}{\partial \kappa} \quad ; \qquad (12)$$

So what we really need are the derivatives of (2) with respect to $\underline{a}$, $\underline{c}$, and $\underline{d}^*$ (substitute all for κ above). To write these, a useful bit of

shorthand is

$$e_k = \exp(a_k\theta + c_k) \quad ,$$

which is always assumed to be evaluated at the appropriate value of θ; all summations of e_k are over categories $k=0,1,\ldots m_j$. The elements of the gradient (P_h is short for $P_j(h;\theta_q)$ hereafter) to be substituted in (12), then (11), and finally zeroed to maximize (10) are:

$$\frac{\partial P_h}{\partial a_0} = \frac{[(\Sigma e_k)\theta d_h e_0 - (e_h + d_h e_0)\theta e_0]}{(\Sigma e_k)^2}$$

$$\frac{\partial P_h}{\partial a_h} = \frac{[(\Sigma e_k)\theta e_h - (e_h + d_h e_0)\theta e_h]}{(\Sigma e_k)^2}$$

$$\frac{\partial P_h}{\partial a_{h'}} = \frac{-\theta e_{h'}}{\Sigma e_k}$$

$$\frac{\partial P_h}{\partial c_0} = \frac{[(\Sigma e_k)d_h e_0 - (e_h + d_h e_0)e_0]}{(\Sigma e_k)^2}$$

$$\frac{\partial P_h}{\partial c_h} = \frac{[(\Sigma e_k)e_h - (e_h + d_h e_0)e_h]}{(\Sigma e_k)^2}$$

$$\frac{\partial P_h}{\partial c_{h'}} = \frac{-e_{h'}}{\Sigma e_k}$$

$$\frac{\partial P_h}{\partial d_h^*} = \frac{e_1}{\Sigma e_k} \cdot \frac{\Sigma d_k^* - d_h^{*2}}{(\Sigma d_k^*)^2}$$

The gradient does not simplify very much since the model is not part of an exponential family. However, it is possible to locate the maximum of ℓ_j using these derivatives, and use the resulting set of parameters in the next E-step. Our estimation procedure allows equality constraints to be placed across as well as within items; this requires some of the gradients to be accumulated across items, but causes no serious problem beyond book-keeping. We use a conditioned Newton-type algorithm (MINIM, described by Haberman, 1974) to locate the maximum of (10); the conditioning is useful since the matrix of second derivatives may be nearly singular for some items.

Convergence and Local Minima

Using the updated item parameters from the M-step, the sequence [E-step, M-step] is repeated until either a) the parameters stabilize or b) a fixed number of cycles is reached (we usually use 15 or 20). The algorithm sometimes converges in the parameter space. More often, it does not, but the lack of convergence causes no real problems. In such cases, most of the parameters (and the loglikelihood) remain roughly constant after ten or fifteen cycles, while a parameter (or a few) change linearly and indefinitely. The changing parameter is frequently a particular α_k (and its associated γ_k), changing as the associated a_k rises toward a very high (possibly infinite) MLE. The difference in goodness of fit with such a slope high (3. to 5. or so) and

much higher appears to be negligible, so there is no loss in simply stopping the estimation procedure arbitrarily. In such a case, the trace line is fairly well-determined; but the numerical value of the associated a_k is not.

A second case in which the parameters fail to clearly converge arises with γ_k's associated with a_k's near zero. The parameter c_k becomes ill-defined when a_k is zero, and the associated γ_k "wanders." Again, there is no loss of fit if the estimation procedure is stopped. The Bock-Aitkin algorithm may be speeded slightly using acceleration as in Thissen (1982), but the value of the acceleration parameter should be limited to about -1, which doubles the step-size at each cycle, to avoid oscillation.

The likelihood surface for the model clearly has more then one "local" minimum. The indeterminacy with respect to reflection of θ gives two equal, identical modes. With poorly chosen starting values, the estimation procedure has located other (apparent) stationary points; usually they seem to be located in a peculiar region corresponding to some items using both reflections of θ in the same solution -- these may be modes or saddle points. Good starting values and the EM-like nature of the Bock-Aitkin algorithm provide a solution in this case, because EM-algorithms only climb local modes. If the algorithm starts near the desired part of the likelihood surface, it will end there. It appears that sensible starting values solve the problem.

The likelihood surface is multi-modal and may be ill-behaved, producing parameter estimates that are sometimes undefined or are on the

boundary. The entire system begs for full Bayesian treatment, with a prior distribution restricting the parameters to a reasonable part of the space under all circumstances. At this stage however, it is not clear what sort of prior may be appropriate. After some experience with the model is gained beyond that in the examples which follow, we may be in a position to specify reasonable parameters for a prior.

Alternatively, a different model, in the exponential family and with a unimodal likelihood (preferably with similar properties to this one in terms of fitting the data), would also solve all of these problems. Such a model has yet to be proposed.

Characterizing $\underline{\theta}$

The parameter estimation described above usually has the goal of "calibrating" a set of test items, after which the parameters are to be taken as known and used to characterize θ for examinees who produce a particular response pattern $\underline{x}$. Given a set of item parameters, the posterior density for θ is

$$P(\underline{x};\theta) = \prod_{j=1,n} P_j(x_j)\phi(\theta) \quad , \qquad (13)$$

in which $P_j(x_j)$ is from equation (2). If the model is correct, (13) describes the distribution of examinees who respond with pattern $\underline{x}$. It can be characterized graphically, and two examples will be presented below in Figure 5.

For more than a few items, (13) is roughly Gaussian in shape, and so it may also be described by estimates of its location and spread. An

extension of the procedure commonly used in binary IRT is to use the mode as an estimate of the location of (13), where

$$\frac{\partial \log P(\underline{x};\theta)}{\partial \theta} = \sum_{j=1,n} \frac{\partial \log P(x_j;\theta)}{\partial \theta} + \frac{\partial \log \phi(\theta)}{\partial \theta}$$

is zero, with the variance approximated by the negative inverse of

$$E \frac{\partial^2 \log P(\underline{x};\theta)}{\partial \theta^2} .$$

The modal estimate is practical to compute as long as $\phi(\theta)$ is a reasonable function, and easy if $\phi(\theta)$ is normal. There is no guarantee that (13) is unimodal; for small numbers of items it is likely to have more than one mode. Multimodality presents potential problems for mechanical use of modal estimates.

However, it is also fairly straightforward to numerically integrate (13) to obtain its mean and variance. The mean has been called the "EAP" (Expected A Posteriori) estimate of θ by Bock and Mislevy (1982). An advantage of the EAP procedure over modal estimation is that the derivative of $\phi(\theta)$ is not required; therefore $\phi(\theta)$ may take any form describable as a histogram with finite variance.

Examples

To illustrate application of the model to item response data, we have analyzed several subsets of items from the ASVAB Form 8A administered to a national probability sample of youth by the National Opinion Research Center; the data are described by Bock and Mislevy

(1981). The data used here are from a 10% subsample of the NORC sample, total N=1178. For small illustrations of the performance of the model, we use four-item subsets consisting of the last four items of the Word Knowledge (WK) and General Science (GS) sections of the test. Four items were used because, with roughly a thousand examinees, that is the largest number for which the number of cells in the m^n table is less than N (4^4=256), so significance tests of the overall goodness of fit of the model against a general multinomial alternative are feasible. The last items were used because they are most difficult, giving the trace lines more latitude to be fitted.

For the four-item examples, only examinees with complete data (no non-response) for all four items were used. For GS, N=1048 and 214 of the 256 cells of the table were filled (one of the 4 x 4 marginal tables of this table makes up Table 1); for WK, N=976 and only 156 response patterns were observed.

We also fitted several forms of the model to the entire set of 35 WK items, and a subset of 12 of the 25 GS items called "physical science" hereafter. (The ASVAB "General Science" section consists of about half biological and half physical science questions; we selected a subset more likely to be unidimensional.) In these analyses, all examinees were used (N=1178) and non-response data to individual items were ignored (not placed in any category).

Results

Table 2 gives summary goodness of fit results for the four-item examples.

Insert Table 2 About Here

Bock's (1972) model and Samejima's (1979) modification are both rejected at the $p<0.05$ level for both sets of data. The model called "ABCD" is a form of (2) in which the vector $\underline{\delta}$ is estimated, but restricted to be equal across items: this represents the hypothesis that a (constant) proportion of those who "don't know" select alternative A, a different proportion B, and so on; but the distribution over A, B, C, and D (regardless of the "correct" response alternative) is the same for all items. For the WK data the model is a great improvement over Samejima's $\underline{\delta}=\underline{0}$ model ($G^2=12.6$ on 3 d.f.), and for GS it is not. The estimated vector $\underline{d}$ for WK for [A,B,C,D] is [.1,.2,.4,.3].

In the last four WK items, D is correct for the first and the last and C is correct for the others; so to test the hypothesis that (somehow) the correct alternative "attracts" guessing we tested "ABCD(C),ABCD(D)" in which the two pairs of items had different vectors $\underline{\delta}$. This model fits the WK four-item data, and the improvement over ABCD is significant ($G^2=10.6$, d.f.=3, $p<0.02$); the parameter estimates are given in table 3. The "saturated" version of (2), with 11 free

parameters per item, does not fit the WK data significantly better.

Insert Tables 3 and 4 About Here

The last four GS items do not have such a convenient structure of correct alternatives; but for these items "correctness" seemed more important than alternative position, so in the model "correct vs. incorrect" d_h is constrained to be one value (the same for all items) for the correct response and another for all the incorrect responses. This model (barely) fits the GS data. The parameter estimates are given in Table 4. Forty percent of the DK guess the correct alternative; there is no evidence that this varies across items (The last line of Table 2 gives the G^2 for "correct vs. incorrect each item").

When the complete WK data are fitted, the likelihood ratio test against a "general multinomial alternative" for the 1178 observations in a 4^{35} table is, of course, meaningless. Nevertheless, we report likelihood ratio tests between variously constrained versions of the model. The Samejima (1979) model is the most constrained form that we use for comparison. A model of the "ABCD" form of (2), in which $\underline{\delta}$ is estimated, equal across items, gives a reduction $G^2=41$ on 3 d.f., $p<0.001$; the estimate of $\underline{d}$ was [.2,.3,.3,.2] for [A,B,C,D]. As with the four-item WK, a model in which $\underline{\delta}$ is constrained to be equal only among items with the same (letter) correct response fits significantly better: $G^2=72$ on 9 d.f., $p<0.01$. No more complex models were considered, as estimation of 11 parameters per item for the very easy items (to which

the majority of examinees responded correctly) would be estimating many parameters with little data. The estimated parameters for the last four items obtained in the analysis of the entire WK set are given directly below their values for the four-item estimation in Table 3.

Analysis of the physical science portion of the GS section gave similar results: (2) in its "ABCD" form fitted significantly better than Samejima's model, $G^2=29$ on 3 d.f., $p<0.01$. The estimate of $\underline{d}$ was again [.2,.3,.5,.2]. And the model in which $\underline{d}$ varied according to the letter-value of the correct alternative was better: $G^2 = 26$ on 9 d.f., $p<0.01$. Two of the items in the "physical science" set are among the last four GS items, and their parameter estimates are given in Table 4.

Discussion

In cases in which the fit can be tested -- with a thousand examinees and four-alternative items, four items -- some forms of the model given by (2) and (5) fit item response data satisfactorily. This represents a major step forward in item analysis, because it is no longer necessary to look at deviations from the fit to examine items; we may examine the fit itself.

Insert Figure 1 About Here

The solid curves in Figure 1 illustrate the trace lines for the "physical science" fit to GS item 23, which is item 2 in Table 1, discussed above. From Table 1 we inferred that those who chose A or B

might be more able than those who chose D. Indeed, the trace line for D restricts that response to those with values of $\theta<0$ for all practical purposes, while B-responses come mostly from those with $\theta>0$, and A is spread all over. The DK trace line indicates that most of those below $\theta=-.5$ have no idea; and more of them guess correctly than any other way.

The dashed curves in Figure 1 illustrate the trace lines for the "four-item" fit to item 23. Note in Table 4 that the parameter estimates differ a great deal, but the solid and dashed trace lines in Figure 1 differ much less (except for DK) in the middle. DK is a completely inferred latent response and is not really very well-defined by four items. The other curves are more similar in the middle, where the data are; 95% of the population distribution lies between θs of -2 and +2. The different parameter estimates seem to affect the curves primarily at the extremes.

Unlike algorthms which make use of point-estimates of θ in the estimation of item parameters, and require both large numbers of examinees and large numbers of items, the MML estimation procedure used here should be consistent considering the number of examinees alone if the model is correct. That is, as the number of examinees becomes large the estimates for four items or 12 or 35 should all converge to the true values. So the dashed and solid lines in Figure 1 should be the same. They are very similar, and there are three possible reasons for the small differences observed between them. First, one thousand examinees may easily not be "asymptotic" for a table with 256 cells. Second, one or the other solution may not be completely converged; EM-algorithms can

be slow near the maximum. Third, and most likely, the model may be incorrect in that the latent dimension θ may be defined slightly differently by the most difficult four items than by the entire test.

Insert Figure 2 About Here

Figure 2 similarly illustrates the four-item and complete test estimates of the trace lines for WK 32, which demonstrates this effect more graphically. With estimates from the entire test, DK goes to unity for low θ. But the four-item (dashed) curve for DK rises for moderately low θ and then goes back down, and A goes to unity for low θ. That can't be right; but it is all happening in the region below θ=-2, where there are essentially no data and the model is extrapolating. Extrapolation is bad: the model is too flexible at the extremes. Flexibility was one of the goals of the model, but it seems we may have overdone it. This is in marked contrast to traditional IRT models which go to unity, zero, or some asymptote at the extremes, and never misbehave there. Note that even with all of this strangeness, the other trace lines in Figure 2 are essentially identical between θs of -2 and +2.

Insert Figure 3 About Here

Figure 3 shows the trace lines for the correct responses only for WK items 32-35 from the four item set and the whole test, as well as

traditional 3-PL curves estimated for the entire test (dotted lines) for comparison. If the "tail-wagging" to the left of θ=-2 is ignored, the pairs of curves for the whole test are nearly identical and the four-item curves are somewhat deviant. That may be capitalization on chance in four items -- or it may be that θ differs mildly in the hardest items of the test from its definition in the entire test. None of the curves are extremely different.

Insert Figure 4 About Here

For WK as a whole, the usual result (of Bock, 1972; Thissen, 1976) is obtained with respect to test information: Figure 4 shows the information curves for the multiple category scoring and the binary scoring (3-PL). For θ<0, information from incorrect responses increases total information by about 50%, equivalent to extending the test (for half the examinees) from 35 to 50 items.

Insert Figure 5 About Here

Figure 5 illustrates the process. The solid curves show the trace lines associated with a particular set of responses [B,B,B,B], all incorrect, for the last four WK items. Then the N(0,1) population distribution is plotted with the the product of all five curves, or posterior density, labelled "Total." The mode of that density is traditionally $\hat{\theta}$; it is about -0.7 (s.e.=0.6) in this case. The dashed

curves show the 3-PL trace lines, all incorrect, so they are all monotonic decreasing. That is not very informative; the (dashed) posterior is broader (s.e.=0.7) and closer to zero ($\hat{\theta}$=-0.5) because there is less information so the estimator is "shrunk" more toward the mean of the population distribution. If the model approximates the world reasonably well, the multiple category scoring gives more information about people responding [BBBB] than can 3-PL.

General Discussion

The model proposed here provides the first practical, complete IRT item analysis for multiple choice tests, describing the performance of all of the response alternatives as functions of the trait being measured. The parameters of the model are not readily interpretable for the most part, but graphical presentation of the trace lines and/or their products with the population distribution gives thorough item analysis which has much to recommend it.

This sort of item analysis is sufficiently flexible that "bad" trace lines are fitted. These can then be observed and such items modified or eliminated in the course of empirical test development. That is not possible, in general, with simpler models which require examination of the residuals to find bad items. For item analysis, the model of (2) and (5) is clearly an excellent choice; for scoring tests (aka "estimating $\hat{\theta}$"), matters are more complicated.

Non-monotonic trace lines for the correct response have become a popular feature of recent IRT developments: Lord (1982) and Choppin

(1983) have proposed item response models which permit "low-ability" non-monotonicity, in work on binary models independent from Samejima's (1979) multiple category proposal. There are a number of possible explanations for non-monotonicity on the left, as for the correct response in Figure 1. The only requirement for candidacy as an explanation for the effect is that it must account for "getting the right answer for the wrong reason." Bock (personal communication, 1983) suggests a name for the phenomenon: "positive misinformation."

Two sources of positive misinformation come to mind. The first is that the correct response for a particular item differs from the distractors on dimensions other than that which is intended and observable, given sufficient ability. Examinees of medium and high ability perceive the features and attempt the processing intended by the item-writer, while examinees of low ability see other features of the alternatives which cause them to select the correct one. These "other features" may be effectively "invisible" to individuals of higher ability, and therefore to the item-writers as well.

A second possible cause of non-monotonicity in the correct trace line is cheating. If low-ability individuals cheat (e.g., by copying a neighbor's answer, which is more likely to be correct than not for most items), then the resulting correct alternative trace lines will rise for those of very low ability.

It may also be possible that non-monotonic trace lines are estimation artifacts in small sets of items (like four). It may be that, with few monotonic correct-alternative trace lines to "orient" θ, the estima-

tion procedure could "wrap around" and (effectively) place some of the high ability individuals on the left end of the θ-continuum. But this explanation can be discounted with longer tests, as the non-monotonicity there comes from correct response to an item being more likely from examinees who get most of the other items wrong than from those who get fewer of the other items wrong.

Distinguishing among these possibilities would require experimentation with the items and the testing situation. The model does, however, offer an item analysis which permits such phenomena to exhibit themselves in the fit, when they are present.

So the use of non-monotonic curves in item analysis may be required to fit the observed data. However, the use of non-monotonic trace lines in constructing the posterior density, a measure of the central tendency of which is to be called $\hat{\theta}$ and used to "score the test", gives rise to many potential problems. If the correct response trace line for item j is non-monotonic on the right (e.g., it turns down), then examinees with some response patterns on the other items will be "penalized" by responding correctly to item j; they would have been assigned a higher $\hat{\theta}$ if they had selected certain of the incorrect alternatives. Correct response trace lines which are non-monotonic on the left similarly "penalize" examinees of low ability who respond correctly. As measurement, this is all probably satisfactory: conditional on the other item responses, a correct response to item j may not imply higher ability; it may be more likely to be guessing or cheating. But this may be a problem for the test "as contest." Further, it might be difficult

to defend such a method of test scoring against the onslaught of members of the bar before a jury. A case can, therefore, be made for either a) rejecting the model or b) rejecting the use of items with non-monotonic trace lines for the correct response.

Multiple category scoring of this nonlinear sort has certain bizarre properties even if non-monotonic trace lines are eliminated. For certain regions on the θ-continuum, selecting a particular incorrect response will increase $\hat{\theta}$ more than would selecting the actual correct response. For instance, in the Word Knowledge item in Figure 2, for θ just below -1, selecting response B will increase $\hat{\theta}$ more than would selecting the correct response. But for θ around 1, selecting B will "penalize" the respondent! The contingencies are sufficiently complex that it is unlikely that the examinees could find a strategy to take advantage of the system. But the possibilities for legal difficulties are considerable under circumstances in which the test and scoring system must be disclosed. On the other hand, the quality of measurement for research purposes and in non-disclosed tests should be improved by these "bizarre" features; that is, after all, where the multiple category model obtains its additional information.

Conclusion

Item response models for multiple choice items have come of age: they fit the data. Questions remain about uses for these models. They produce excellent item analysis, but it is complex and best-represented graphically -- that breaks with the tradition of item analysis with a

few numerical summaries. Multiple category scoring clearly increases the information obtained in scoring a test -- but again at the cost of complexity, and, potentially, controversy. The validity of test scored with such methods has not been examined here, nor can it be with internal consistency data in any event. Consideration of all of these matters is deferred to future work.

References

Bock, R.D. Estimating item parameters and latent ability when the responses are scored in two or more nominal categories. Psychometrika, 1972, 37, 29-51.

Bock, R.D. and Aitkin, M. Marginal maximum likelihood estimation of item parameters: An application of an EM algorithm. Psychometrika, 1981, 46, 443-459.

Bock, R.D. and Lieberman, M. Fitting a response model for n dichotomously scored items. Psychometrika, 1970, 35, 179-197.

Bock, R.D. and Mislevy, R.G. Data quality analysis of the Armed Forces Vocational Aptitude Battery. DOD MDA 903-802-0333. National Opinion Research Center, Chicago, 1981.

Bock, R.D. and Mislevy, R.G. Applications of EAP estimation in computerized adaptive testing. Paper presented at the annual meeting of the Psychometric Society, Montreal, 1982.

Choppin, B. A two-parameter latent trait model. CSE Report no. 197, Center for the Study of Evaluation, Graduate School of Education, University of California, Los Angeles, 1983.

Dempster, A.P., Laird, N.M., and Rubin, D.B. Maximum likelihood from incomplete data via the EM algorithm (with Discussion). Journal of the Royal Statistical Society, Series B, 1977, 39, 1-38.

Haberman, S. Subroutine MINIM. In R. D. Bock and Bruno Repp (Eds.) MATCAL: Double precision matrix operations subroutines for theIBM System 360/370 computers. Chicago: National Educational Resources, 1974.

Levine, M.V. and Drasgow, F. The relation between incorrect option choice and estimated ability. Educational and Psychological Measurement, 1982, in press.

Lord, F.M. Maximum likelihood estimation of item response parameters when some responses are omitted. In preparation, 1982.

Samejima, F. A new family of models for the multiple choice item. Research Report #79-4, Department of Psychology, University of Tennessee, 1979.

Sympson, J.B. A new IRT model for calibrating multiple choice items. Paper presented at the annual meeting of the Psychometric Society, Los Angeles, 1983.

Thissen, D. Information in wrong responses to the Raven Progressive Matrices. Journal of Educational Measurement, 1976, 13, 201-214.

Thissen, D. Marginal maximum likelihood estimation for the one-parameter logistic model. Psychometrika, 1982, 47, 175-186.

Footnotes

This research was supported in part by ONR contract N00014-83-C-0283 to the University of Chicago, R. Darrell Bock, Principal Investigator and AFHRL contract F41689-82-C-00020 to the Educational Testing Service, Howard Wainer, Principal Investigator. This project was facilitated by a sabbatical leave provided DT by the University of Kansas, during which time some of this work was done with the Research Statistics Group of the Educational Testing Service. Conversations with Darrell Bock, Howard Wainer, Paul Holland, Frederic Lord, Donald Rubin and James Ramsay have also been important and useful in shaping the research and its presentation here. While all of those mentioned have helped us gain wisdom in these matters, any error that remains is, of course, our own.

Table 1.

Cross classification of responses to two items (22 and 23) of the ASVAB Form 8A General Science subtest for 1048 examinees of the NORC national probability sample (Bock and Mislevy, 1981).

		Item 2				
		A	B	C*	D	
	A	20(22)	107(23)	79(22)	44(31)	250
Item	B*	32(36)	157(34)	143(40)	35(25)	367
1	C	7(8)	36(8)	24(7)	19(14)	86
	D	30(33)	161(35)	112(31)	42(30)	345
		89	461	358	140	1048

*=Correct;

Column percentages in parentheses.

Table 2.

G^2 values for selected models for the four-item examples.

		Word Knowledge		General Science	
Model	d.f.	G^2	p	G^2	p
Bock (1972)	231	307	<.01	277	<.02
Samejima (1979)	223	271	<.02	264	<.05
"ABCD"	220	258	<.04	262	<.02
"ABCD(C),ABCD(D)"	217	248	>.05	---	--
"11 per item"	211	245	<.05	---	--
"Correct vs. incorrect"	222	---	--	253	=.05
"Correct vs., each item"	3	---	--	253	=.05

Table 3.

Estimated parameters for the last four Word Knowledge items.

Estimates for four items above, whole test below.

Parameters for correct response underscored.

		Response				
		DK	A	B	C	D
Item						
32	a_k	-1.1	-2.9	1.5	0.2	2.2
		-1.7	-1.0	1.1	0.3	1.9
	c_k	0.1	-3.8	2.9	-1.9	2.7
		0.3	-2.3	2.4	-2.5	2.1
	d_k		0.1	0.2	0.3	0.4
			0.25	0.25	0.25	0.25
33	a_k	-3.0	-0.3	1.2	2.8	-0.7
		-2.1	-0.6	1.2	2.3	-0.8
	c_k	0.1	0.6	-1.6	0.1	0.8
		2.1	0.5	-3.0	-0.6	1.0
	d_k		0.2	0.2	0.5	0.1
			0.2	0.2	0.4	0.2
34	a_k	-0.8	-1.8	-0.1	1.9	0.7
		-1.3	-0.9	-0.2	1.9	0.5
	c_k	-1.6	-3.2	0.2	2.6	2.0
		-0.9	-2.5	-0.1	1.8	1.6
	d_k		0.2	0.2	0.5	0.1
			0.2	0.2	0.4	0.2
35	a_k	-1.6	-0.4	-0.2	-0.8	3.0
		-1.9	0.5	0.0	-0.6	1.9
	c_k	0.0	-1.3	0.4	0.4	0.5
		-0.1	-2.0	0.5	0.8	0.8
	d_k		0.1	0.2	0.3	0.4
			0.25	0.25	0.25	0.25

Table 4.

Estimated parameters for the last four General Science items.

Estimates for four items above, "physical science" subset below.

Parameters for correct response underscored.

		Response				
		DK	A	B	C	D
Item 22	a_k	-2.3	-0.2	2.0	0.9	-0.3
	c_k	0.5	0.7	-0.5	-1.9	1.1
	d_k		0.2	0.4	0.2	0.2
23	a_k	-0.8	0.6	-0.5	1.1	-0.4
		-4.0	1.2	1.1	2.5	-0.9
	c_k	1.6	-2.8	1.5	0.0	-0.3
		-0.8	-0.7	1.4	0.2	-0.1
	d_k		0.2	0.2	0.4	0.2
			0.2	0.3	0.3	0.2
24	a_k	-0.5	-0.2	2.0	-1.2	0.0
		-3.3	0.2	2.6	-0.1	0.5
	c_k	-0.3	0.7	-1.0	0.7	0.0
		-1.8	1.1	-1.0	1.4	0.2
	d_k		0.2	0.4	0.2	0.2
			0.2	0.4	0.2	0.2
25	a_k	-1.5	-0.7	-0.2	0.1	2.3
	c_k	0.4	0.4	-0.5	0.5	-0.8
	d_k		0.2	0.2	0.2	0.4

Figure Captions

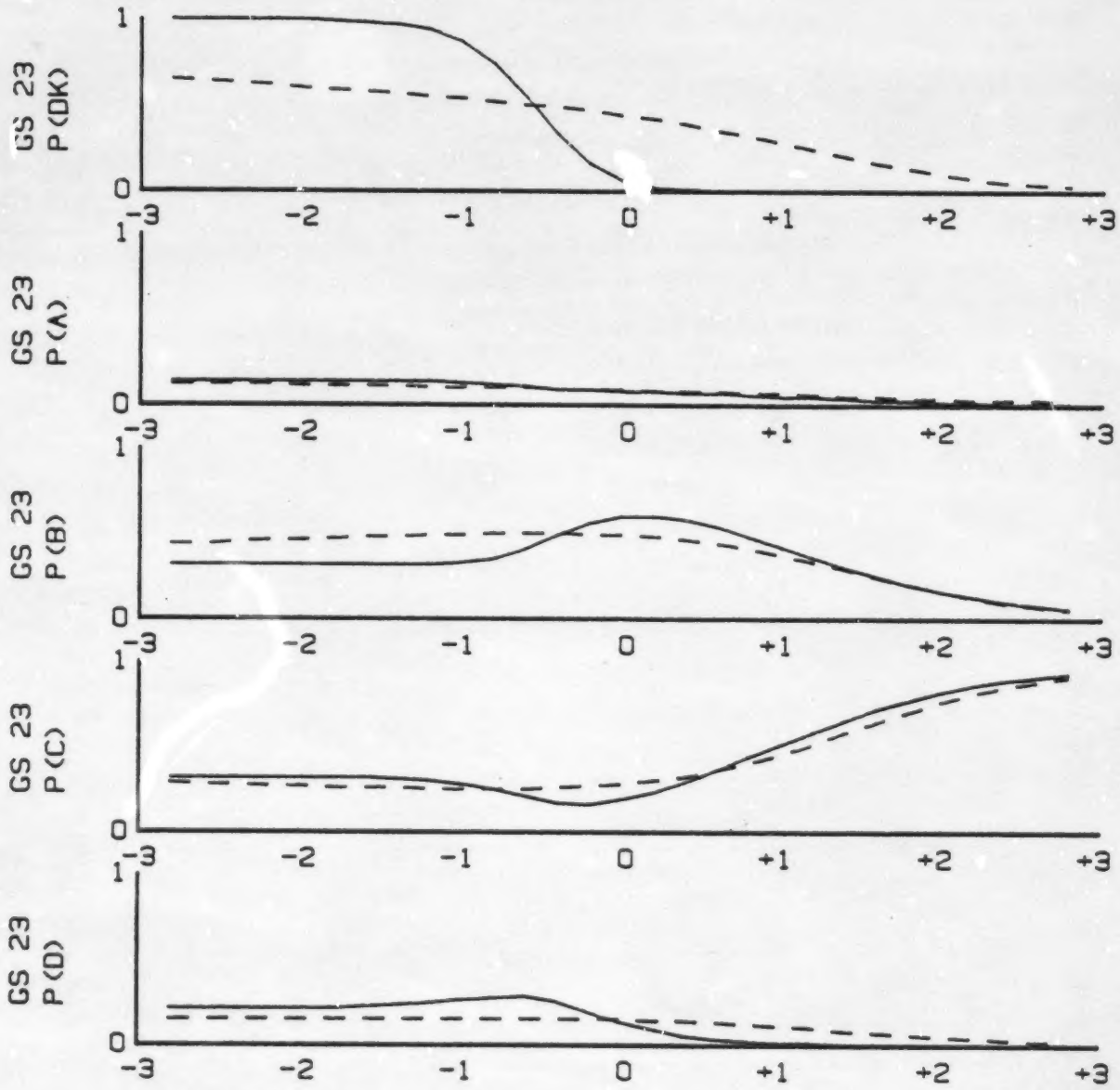
Figure 1. Five panels showing P(k), k=0,1,...,m_j as a function of θ; the solid lines correspond to the probabilities for DK, and responses A,B,C, and D for GS item 23, estimated as part of "physical science." The dashed curves use the estimates obtained with the last four items.

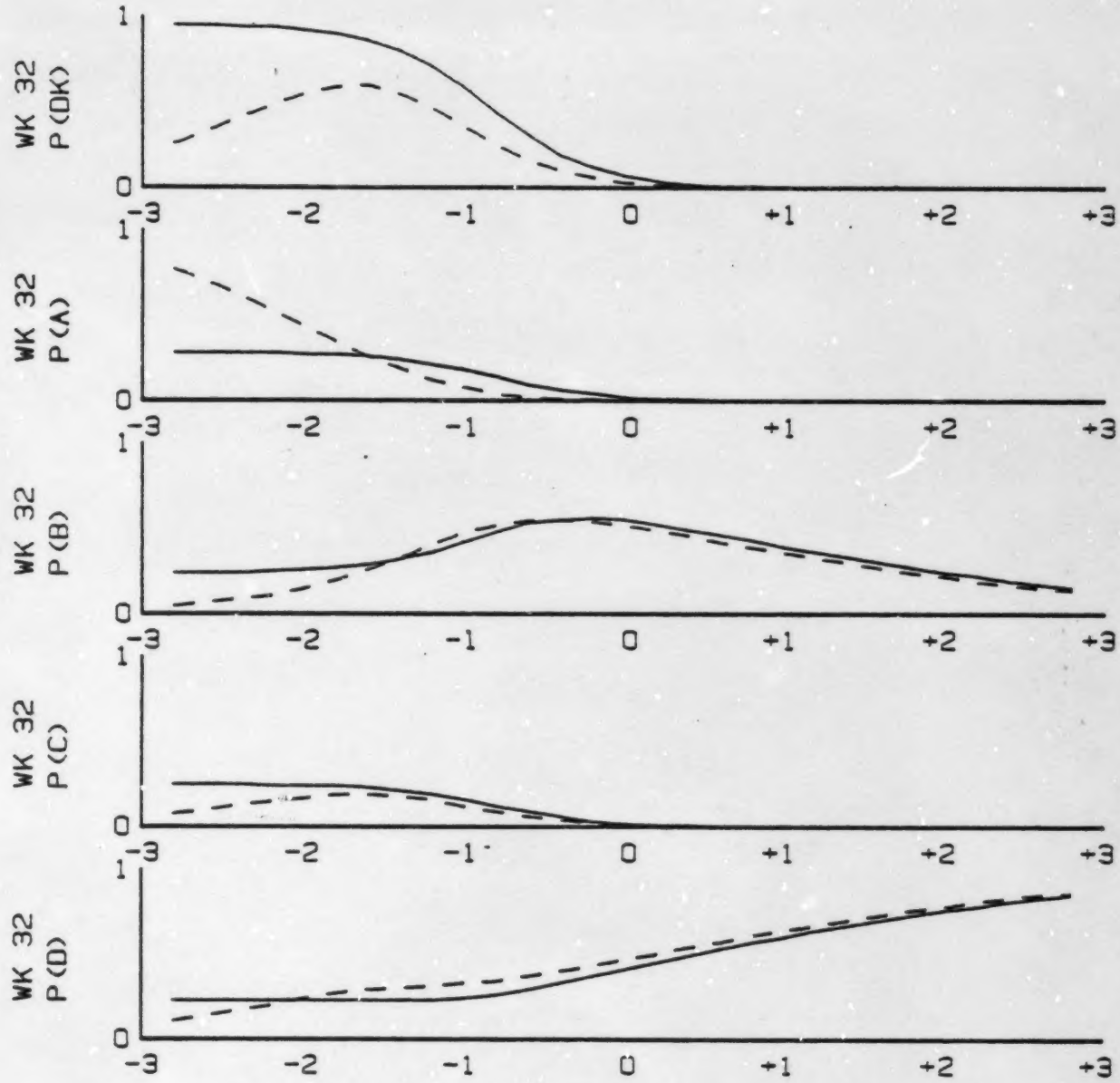
Figure 2. Five panels showing P(k), k=0,1,...,m_j as a function of θ; the solid lines correspond to the probabilities for DK, and responses A,B,C, and D for WK item 32 estimated in the entire test. The dashed curves use the estimates obtained with the last four items.

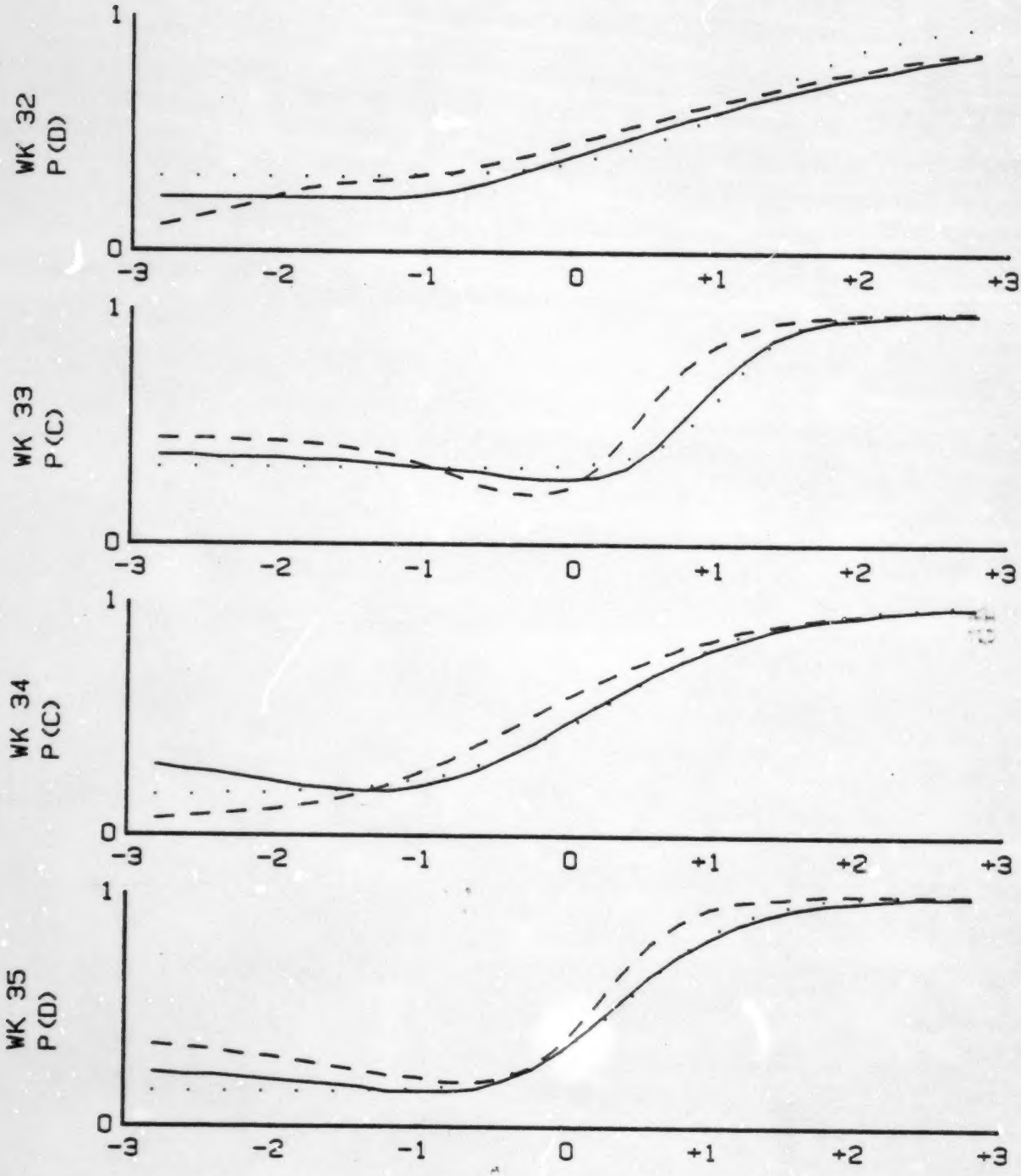
Figure 3. Four panels show three alternative fits of the trace line for the correct responses for the last four WK items: dots are on the standard 3-PL curve, the solid line is the model of the present paper fitted with the entire test, and the dashed line is the present model fitted with only those items.

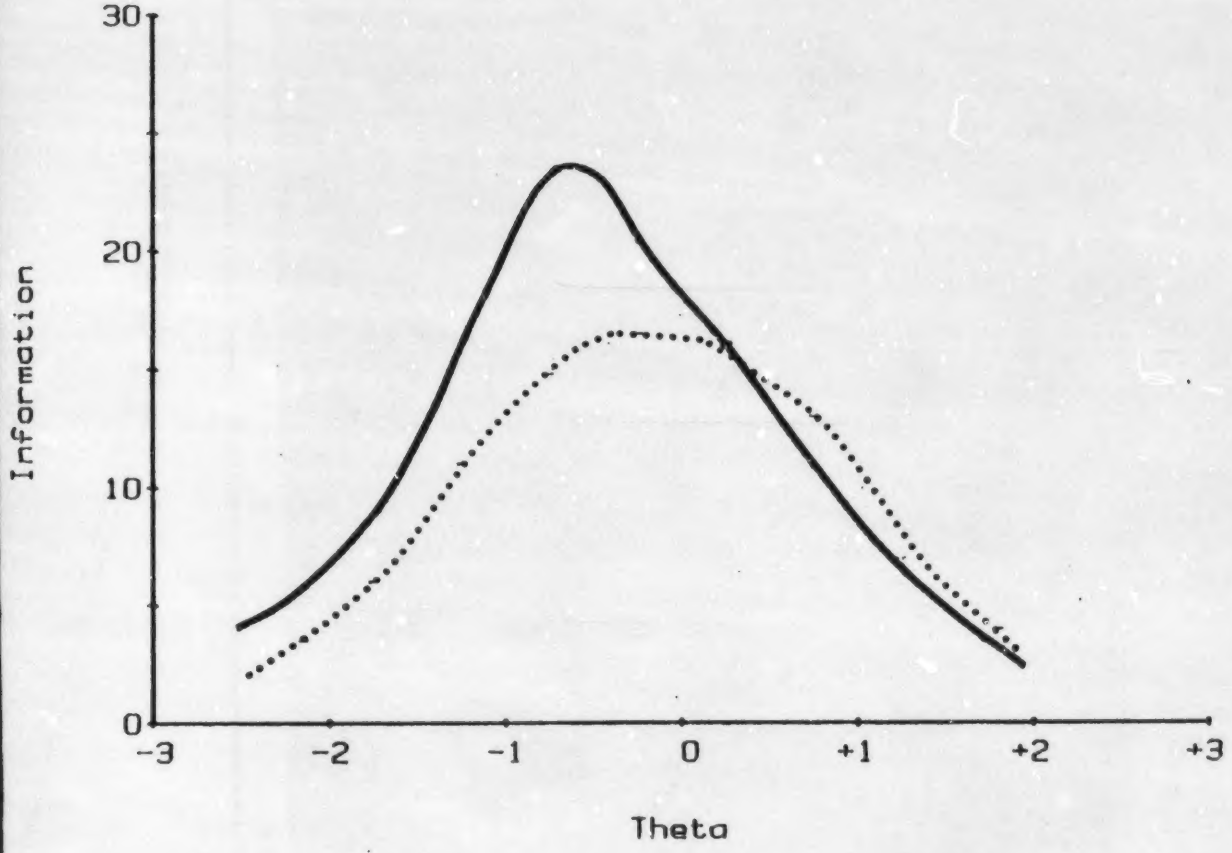
Figure 4. Test information curves for binary (dotted line) and multiple category (solid line) scoring for WK, as functions of θ.

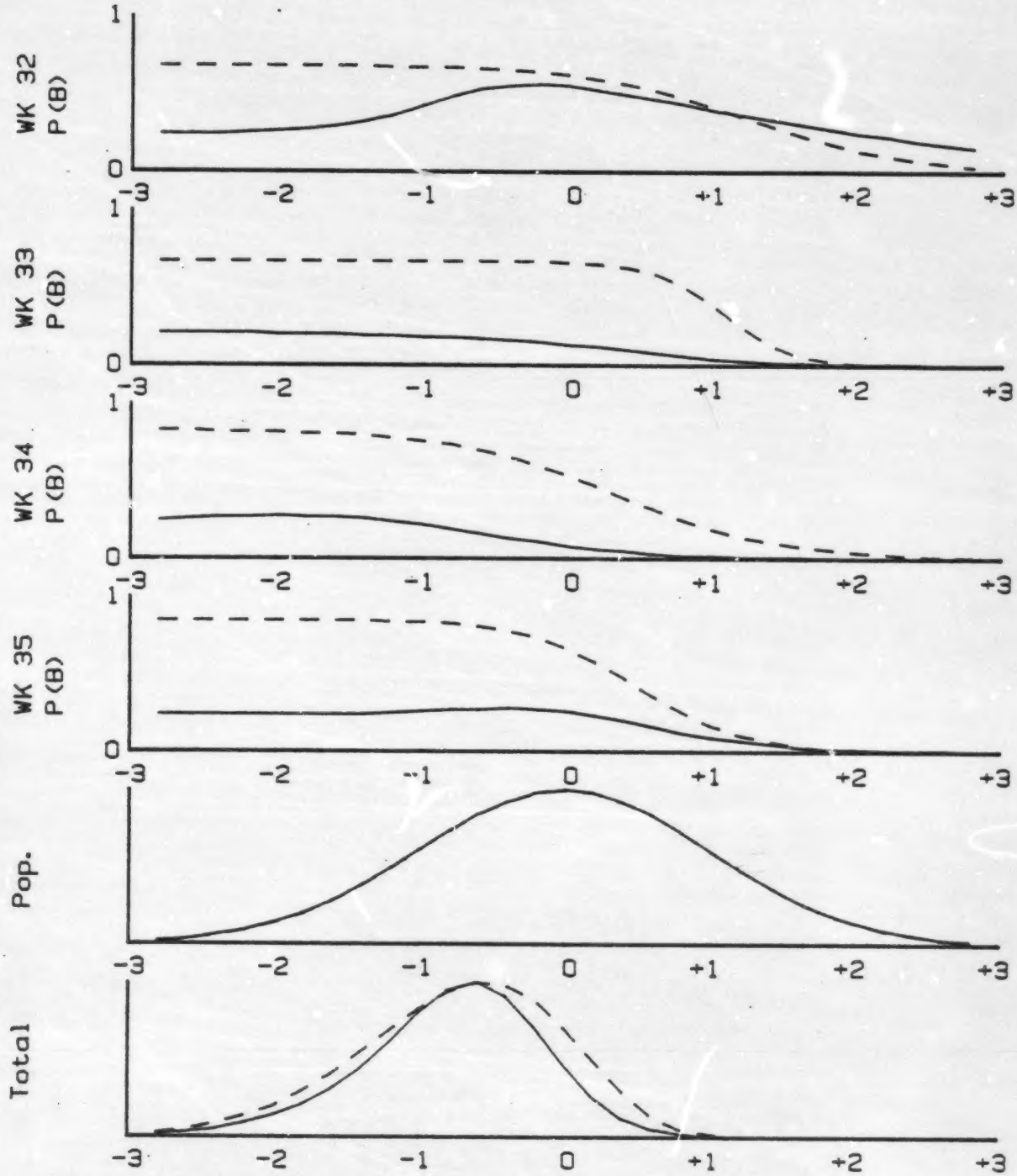
Figure 5. The solid lines give the trace lines from the present model and posterior density (labelled "Total") for response pattern [BBBB] (all incorrect) for the last four items of WK. Dashed lines give the corresponding curves for the same response pattern using 3-PL estimates.


GS 23
P(DK)
GS 23
P(A)
GS 23
P(B)
GS 23
P(C)
GS 23
P(D)
1
0
-3
-2
-1
0
+1
+2
+3


WK 32
P(OK)
WK 32
P(A)
WK 32
P(B)
WK 32
P(C)
WK 32
P(D)
1
0
-3
-2
-1
0
+1
+2
+3


WK 32
P(D)
1
0
-3
-2
-1
0
+1
+2
+3
WK 33
P(C)
1
0
-3
-2
-1
0
+1
+2
+3
WK 34
P(C)
1
0
-3
-2
-1
0
+1
+2
+3
WK 35
P(D)
1
0
-3
-2
-1
0
+1
+2
+3


30
20
10
0
Information
-3
-2
-1
0
+1
+2
+3
Theta


WK 32
P(B)
1
0
-3
-2
-1
0
+1
+2
+3
WK 33
P(B)
WK 34
P(B)
WK 35
P(B)
Pop.
Total

Lowell Schoer
Psycho. & Quanti. Foundations
College of Education
Univesity of Iowa
Iowa City, IA 52242

Dr. Edwin Shirkey
Department of Psychology
Univer. of Central Florida
Orlando, FL 32816

Dr. William Sims
Center for Naval Analysis
200 North Beauregard Street
Alexandria, VA 22311

Dr. H. Wallace Sinaiko
Smithsonian Institution
801 North Pitt Street
Alexandria, VA 22314

Dr. Kathryn T. Spoehr
Psychology Department
Brown University
Providence, RI 02912

Dr. Robert Sternberg
Department of Psychology
Yale University
Box 11A, Yale Station
New Haven, CT 06520

Dr. Peter Stoloff
Center for Naval Analysis
200 North Beauregard Street
Alexandria, VA 22311

Dr. William Stout
University of Illinois
Department of Mathematics
Urbana, IL 61801

Dr. Patrick Suppes
Insti. for Mathematical Studies
in the Social Sciences
Stanford University
Stanford, CA 94305

Dr. Hariharan Swaminathan
Lab. of Psy. & Eval. Research
School of Education
University of Massachusetts
Amherst, MA 01003

Dr. Kikumi Tatsuoka
Comp. Based Educ. Research
252 Eng. Research Laboratories
Urbana, IL 61820

Dr. Maurice Tatsuoka
220 Education Building
1310 South Sixth Street
Champaign, IL 61820

Dr. David Thissen
Department of Psychology
University of Kansas
Lawrence, KS 66044

Dr. Douglas Towne
Univer. of So. California
Behavioral Technology Labs
1845 S. Elena Avenue
Redondo Beach, CA 90277

Dr. Robert Tsutakawa
Department of Statistics
University of Missouri
Columbia, MO 65201

Dr. J. Uhlaner
Uhlaner Consultants
4258 Bonavita Drive
Encino, CA 91436

Dr. V. R. R. Uppuluri
Union Carbide Corporation
Nuclear Division
P.O. Box Y
Oak Ridge, TN 37830

Dr. David Vale
Assessment Systems Corporation
2233 University Avenue
Suite 310
St. Paul, MN 55114

Dr. Kurt Van Lehn
Xerox PARC
3333 Coyote Hill Road
Palo Alto, CA 94304

Dr. Howard Wainer
Division of Psy. Studies
Educational Testing Service
Princeton, NJ 08540

Dr. Michael T. WAller
Department of Educ. Psychology
Univeristy of Wis.-Milwaukee
Milwaukee, WI 53201

Dr. Brian Waters
HumRRO
300 North Washington
Alexandria, VA 22314

Dr. Phyllis Weaver
2979 Alexis Drive
Palo Alto, CA 94304

Dr. David J. Weiss
N660 Elliott Hall
University of Minnesota
75 E. River Road
Minneapolis, MN 55455

Dr. Donald O. Weitzman
Mitre Corporation
1820 Dolley Madison Blvd
McLean, VA 22102

Dr. Rand R. Wilcox
University of Southern Califor
Department of Psychology
Los Angeles, CA 90007

Wolfgang Wildgrube
Streitkraefteamt
Box 20 50 03
D-5300 Bonn 2
WEST GERMANY

Dr. Bruce Williams
Depart. of Educ. Psychology
University of Illinois
Urbana, IL 61801

Dr. Wendy Yen
CTB/McGraw Hill
Del Monte Research Park
Monterey, CA 93940

H. William Greenup
Education Advisor (E031
Education Center MCDEC
Quantico, VA 22134

Dr. Douglas H. Jones
Advanced Statistical Technologies Corporation
10 Trafalgar Court
Lawrenceville, NJ 08148

Professor John A. Keats
Department of Psychology
The University of Newcastle
N.S.W. 2308
AUSTRALIA

Dr. Marcy Lansman
University of North Carolina
Davie Hall 013A
Capel Hill, NC 27514

Dr. Alan Lesgold
Learning R&D Center
University of Pittsburgh
3939 O'Hara Street
Pittsburgh, PA 15260

Dr. Michael Levine
Depart. of Educa. Psychology
210 Education Bldg.
University of Illinois
Champaign, IL 61801

Dr. Charles Lewis
Fac. Soc. Wetenschappe
Rijksuniversiteit Groningen
Oude Boteringestraat 23
9712GC Groningen, NETHERLANDS

Dr. Robert Linn
College of Education
University of Illinois
Urbana, IL 61801

Mr. Phillip Livingston
Systems & Applied Sci. Corp.
6811 Kenilworth Avenue
Riverdale, MD 20840

Dr. Robert Lockman
Center for Naval Analysis
200 North Beauregard Street
Alaexandria, VA 22311

Dr. Frederic M. Lord
Educational Testing Service
Princeton, NJ 08541

Dr. James Lumsden
Department of Psychology
University of Western Australia
Nedlands W.A. 6009
AUSTRALIA

Dr. Gary Marco
Stop 31-E
Educational Testing Service
Princeton, NJ 08451

Dr. Scott Maxwell
Department of Psychology
University of Notre Dame
Notre Dame, IN 46556

Dr. Samuel T. Mayo
Loyola University of Chicago
820 North Michigan Avenue
Chicago, IL 60611

Mr. Robert McKinley
American College Testing Pro.
P.O. Box 168
Iowa City, IA 52243

Dr. Barbara Means
Human Resources Research Organ.
300 North Washington
Alexandria, VA 22314

Dr. Allen Munro
Behavioral Tech. Laboratories
1845 Elena Ave., Fourth Floor
Redondo Beach, CA 90277

Dr. W. Alan Nicewander
University of Oklahoma
Department of Psychology
Oklahoma City, OK 73069

Dr. Donald A. Norman
Cognitive Science, C-015
Univ. of California, San Diego
La Jolla, CA 92093

Dr. Melvin R. Novick
356 Lindquist Cnt. for Measure.
University of Iowa
Iowa City, IA 52242

Dr. James Olson
WICAT, Inc.
1875 Wouth State Street
Orem, UT 84057

Wayne M. Patience
American Council on Education
GED Testing Service, Suite 20
One Dupon Circle, NW
Washington, DC 20036

Dr. James A. Paulson
Portland State University
P.O. Box 751
Portland, OR 97207

Dr. James W. Pellegrino
University of California,
Santa Barbara
Department of Psychology
Santa Barabara, CA 93106

Dr. Mark D. Reckase
ACT
P.O. Box 168
Iowa City, IA 52243

Dr. Lauren Resnick
LRDC
University of Pittsburgh
3939 O'Hara Street
Pittsburg, PA 15260

Dr. Thomas Reynolds
University of Texas-Dallas
Marketing Department
P.O. Box 688
Richardson, TX 75080

Dr. Andrew M. Rose
Amer. Insti. for Research
1055 Thomas Jefferson St. NW
Washington, DC 20007

Dr. Lawrence Rudner
403 Elm Avenue
Takoma Park, MD 20012

Dr. J. Ryan
Department of Education
Univeristy of South Carolina
Columbia, SC 29208

Prof. Fumiko Samejima
Department of Psychology
Univeristy of Tennessee
Knoxville, TN 37916

Dr. Walter Schneider
Psychology Department
603 E. Daniel
Champaign, IL 61820

Dr. Kazuo Shigemasu
7-9-24 Kugenuma-Kaigan
Fujusawa 251
JAPAN

Dr. Erling B. Andersen
Department of Statistics
Studiestraede 6
1455 Copenhagen
DENMARK

Dr. Isaac Bejar
Educational Testing Service
Princeton, NJ 08450

Dr. Menucha Birenbaum
School of Education
Tell Aviv University
Tel Aviv, Ramt Aviv 60078
ISRAEL

Dr. Werner Birke
Personalstammamt der Bundeswehr
D-5000 Hoeln 90
WEST GERMANY

Dr. Robert Brennan
American College Testing Programs
P.O. Box 168
Iowa City, IA 52243

Dr. Glenn Bryan
6208 Poe Road
Bethesda, MD 20817

Dr. Ernest R. Cadotte
307 Stokely
University of Tennnessee
Knoxville, TN 37916

Dr. John B. Carroll
409 Elliott Road
Chapel Hill, NC 27514

Dr. Norman Cliff
Department of Psychology
Univ. of So. California
University Park
Los Angeles, CA 90007

Dr. Lloyd Humphreys
Department of Psychology
University of Illinois
603 East Daniel Street
Champaign, IL 61820

Dr. William Koch
University of Texas-Austin
Measurement and Evaluation Center
Austin, TX 78703

Dr. Allan M. Collins
Bolt Beranek & Newman, Inc.
50 Moulton Street
Cambridge, MA 02138

Dr. Hans Crombag
Education Research Center
University of Leyden
Boerhaavelaan 2
2334 EN Leyden, NETHERLANDS

CTB/McGraw-Hill Library
2500 Garden Road
Monterey, CA 93940

Dr. Dattpradad Divgi
Syracuse University
Department of Psychology
Syracuse, NE 33210

Dr. Fritz Drasgow
Department of Psychology
Univeristy of Illinois
603 E. Daniel St.
Champaign, IL 61820

Dr. Susan Embertson
Psychology Department
University of Kansas
Lawrence, KS 66045

ERIC Facility-Acquisitions
4833 Rugby Avenue
Bethesda, MD 20014

Dr. Benjamin A. Fairbank, Jr.
McFann-Gray & Associates, Inc.
5825 Callaghan
Suite 225
San Antonio, TX 78228

Dr. Leonard Feldt
Lindquist Center for Measurement
Univesity of Iowa
Iowa City, IA 52242

Dr. Earl Hunt
Department of Psychology
University of Washington
Seattle, WA 98105

Dr. Jack Hunter
2122 Coolidge Street
Lansing, MI 48906

Dr. Richard L. Ferguson
The American College Testing Program
P.O. Box 168
Iowa City, IA 52240

Univ. Prof. Dr. Gerhard Fischer
Liebiggasse 5/3
A 1010 Vienna
AUSTRIA

Professor Donald Fitzgerald
University of New England
Armidale, New South Wales 2351
AUSTRALIA

Dr. Dexter Fletcher
University of Oregon
Department of Computer Science
Eugene, OR 97403

Dr. John R. Rfederiksen
Bolt Beranek & Newman
50 Moulton Street
Cambridge, MA 02138

Dr. Janice Gifford
University of Massachusetts
School of Education
Amherst, MA 01002

Dr. Robert Glaser
Learning Research & Dev. Cente
University of Pittsburgh
3939 O'Hara Street
Pittsburgh, PA 15260

Dr. Bert Green
Johns Hopkins University
Department of Psychology
Charles & 34th Street
Baltimore, MD 21218

Dr. Ron Hambleton
School of Education
University of Massachusetts
Amherst, MA 01002

Dr. Paul Horst
677 G Street, #184
Chula Vista, CA 90010

Dr. Huynh Huynh
College of Education
University of South Carolina
Columbia, SC 29208

AIR FORCE HUMAN RESOURCES LAB
AFHRL/MPD
BROOKS AFB, TX 78235

AIR UNIVERSITY LIBRARY
AUL/LSE 76/443
MAXWELL AFB, AL 36112

DR. ALFRED R. FREGLY
AFOSR/NL
BOLLING AFB, DC 20332

MR. RANDOLPH PARK
AFHRL/MOAN
BROOKS AFB, TX 78235

3700 TCHTW/TTGHR
2LT TALLARIGO
SHEPPARD AFB, TX 76311

DR. JOSEPH YASATUKE
AFHRL/LRT
LOWRY AFB, CO 80230

DR. WILLIAM GRAHAM
TESTING DIRECTORATE
MEPCOM/MEPCT-P
FT. SHERIDAN, IL 60037

DR. WAYNE SELLMAN
OFF. OF THE ASS. SEC. OF
DEFENSE (MRA & L)
2B269 the Pentagon
WASHINGTON, DC 20301

DR. ARTHUR MELMED
724 BROWN
US DEPARTMENT OF ED.
WASHINGTON, DC 20208

MR. THOMAS A. WARM
US COAST GUARD INST.
P.O. SUBSTATION 18
OKLAHOMA CITY, OK 73169

Technical Documants Center
AF Human Resources Lab
WPAFB, OH 45433

DR. EARL A. ALLUISI
HQ, AFHRL (AFSC)
BROOKS AFB, TX 78235

DR. GENEVIEVE HADDAD
PROGRAM MANAGER
LIFE SCIENCES DIRECTORATE
AFOSR
BOLLING AFB, DC 20332

DR. ROGER PENNELL
AF HUMAN RESOURCES LAB
LOWRY AFB, CO 80230

LT. COL JAMES E. WATSON
HQ USAF/MPXOA
THE PENTAGON
WASHINGTON, DC 20330

DEFENSE TECHNICAL INFO CEN.
CAMERON STATION, BLDG 5
ALEXANDRIA, VA 22314
ATTN: TC

Jerry Lehnus
HQ MEPCOM
Attn: MEPCT-P
Fort Sheridan, IL 60037

MAJOR JACK THORPE
DARPA
1400 WILSON BLVD.
ARLINGTON, VA 22209

DR. ANDREW R. MOLNAR
OFF OF SCI & ENG PER & ED
NAT'L SCIENCE FOUNDATION
WASHINGTON, DC 20415

DR. JOSEPH L. YOUNG, DIR.
MEMORY & COGNITIVE PROCESSES
NAT'L SCIENCE FOUNDATION
WASHINGTON, DC 20550

US AF OFFICE OF SCI. RES.
LIFE SCI. DIRECTORATE, NL
BOLLING AIR FORCE BASE
WASHINGTON, DC 20332

MR. RAYMOND E. CHRISTAL
AFHRL/MOE
BROOKS AFB, TX 78235

DR. T. M. LONGRIDGE
AFHRL/OTE
WILLIAMS AFB, AZ 85224

DR. MALCOM REE
AFHRL/MP
BROOKS AFB, TX 78235

MAJOR JOHN WELSH
AFHRL/MOAN
BROOKS AFB, TX

DR. CRAIG I. FIELDS
ADVANCED RES. PROJECTS AGENCY
1400 WILSON BLVD.
ARLINGTON, VA 22209

MILITARY ASS. FOR TRAINING &
PERSONNEL TECH.
OFF. OF THE UNDER SEC. FOR RE
ROOM 3D129, the PENTAGON
WASHINGTON, DC 20301

DR. SUSAN CHIPMAN
LEARNING & DEVELOPMENT
NAT'L INST. OF EDUCATION
1200 19th STREET NW
WASHINGTON, DC 20208

DR. VERN W. URRY
PERSONNEL R&D CENTER
OFF OF PERSONNEL MANAGEMENT
1900 E STREET NW
WASHINGTON, DC 20415

DR. HAMES ALGINA
UNIVERSITY OF FLORIDA
GAINESVILLE, FL 326

MR. BRAD SYMPSON
NAVY PERSONNEL R&D CENTER
SAN DIEGO, CA 92152

DR. MARTIN A. TOLCOTT
ASS. DIR. FOR LIFE SCIENCE
800 NORTH QUINCY STREET
ARLINGTON, VA 22217

DR. JAMES TWEEDDALE
TECHNICAL DIRECTOR
NAVY PERSONNEL R&D CENTER
SAN DIEGO, CA 92152

PERSONNEL & TRAINING
RESEARCH GROUP
CODE 442PT
OFFICE OF NAVAL RESEARCH
ARLINGTON, VA 22217

SPECIAL ASST. FOR EDUCATION
AND TRAINING (OP-01E)
RM. 2705 ARLINGTON ANNEX
WASHINGTON, DC 20370

~~CHIEF OF NAVAL~~
OPERATIONS-
~~RESEARCH-DEVELOPMENT-&~~
~~---STUDIES-BRANCH-~~

LT FRANK C. PETHO, MSC,
USN (PH.D)
CNET (N-432)
NAS
PENSACOLA, FL 32508

OFFICE OF THE CHIEF OF NAVAL
OPERATIONS
RES. DEV. & STUDIES BRANCH
OP 115
WASHINGTON, DC 20350

DR. BERNARD RIMLAND (01C)
NAVY PERSONNEL R&D CENTER
SAN DIEGO, CA 95152

DR. CARL ROSS
CNET-PDCD
BUILDING 90
GREAT LAKES NTC, IL 60088

DR. ROBERT G. SMITH
OFFICE OF CHIEF OF NAVAL OP.
OP-987H
WASHINGTON, DC 20350

DR. ALFRED F. SMODE, DIR.
TRAINING ANALYSIS &
EVALUATION GROUP
DEPT. OF THE NAVY
ORLANDO, FL 32813

DR. R. SNOW, LIAISON SCI.
OFFICE OF NAVAL RESEARCH
BRANCH OFFICE, LONDON
BOX 39
FPO NEW YORK, NY 09510

DR. RICHARD SORENSEN
NAVY PERSONNEL R&D CENTER
SAN DIEGO, CA 92152

DR. FREDERICK STEINHEISER
CNO - OP115
NAVY ANNEX
ARLINGTON, VA 20370

DR. FRANK VICINO
NAVY PERSONNEL R&D CENTER
SAN DIEGO, CA 92152

DR. EDWARD WEGMAN
OFFICE OF NAVAL RESEARCH
800 NORTH QUINCY STREET
ARLINGTON, VA 22217

DR. RONALD WEITZMAN
NAVAL POSTGRADUATE SCHOOL
DEPT. OF ADM. SCIENCES
MONTEREY, CA 93940

DR. DOUGLAS WETZEL
CODE 12
NAVY PERSONNEL R&D CENTER
SAN DIEGO, CA 92152

DR. MARTIN F. WISKOFF
NAVY PERSONNEL R&D CENTER
SAN DIEGO, CA 92152

MR. JOHN H. WOLFE
NAVY PERSONNEL R&D CENTER
SAN DIEGO, CA 92152

DR. WALLACE WULFECK, III
NAVY PERSONNEL R&D CENTER
SAN DIEGO, CA 92152

DR. NICK BOND
OFFICE OF NAVAL RESEARCH
LIAISON OFFICE, FAR EAST
APO SAN FRANCISCO, CA 96503

LT. ALEXANDER BORY
APPLIED PSYCHOLOGY
MEASUREMENT DIVISION
NAMRL
NAS PENSACOLA, FL 32508

DR. ROBERT CARROLL
NAVOP 115
WASHINGTON, DC 20370

DR. ROBERT BREAUX
NAVTRAEQUIPCEN
CODE N-095R
ORLANDO, FL 32813

Chief of Naval Ed. & Training
Liason Office
Air Force Human Resource Lab
Operations Training Division
WILLIAMS AFB, AZ 85224

DR. STANLEY COLLYER
OFFICE OF NAVAL TECHNOLOGY
800 N. QUINCY STREET
ARLINGTON, VA 22217

CDR MIKE CURRAN
OFFICE OF NAVAL RESEARCH
800 N. QUINCY ST.
CODE 270
ARLINGTON, VA 22217

MIKE DURMEYER, INSTRUCTIONAL
PROGRAM DEVELOPMENT
BUILDING 90
NET-PDCD
GREAT LAKES NTC, IL 60088

DR. RICHARD ELSTER
DEPT. OF ADM. SCIENCES
NAVAL POSTGRADUATE SCHOOL
MONTEREY, CA 93940

DR. PAT FEDERICO
CODE P13
NPRDC
SAN DIEGO, CA 92152

DR. CATHY FERNANDES
NAVY PERSONNEL R&D CENTER
SAN DIEGO, CA 92152

MR. PAUL FOLEY
NAVY PERSONNEL R&D CENTER
SAN DIEGO, CA 92152

DR. ED HUTCHINS
NAVY PERSONNEL R&D CENTER
SAN DIEGO, CA 92152

DR. WILLIAM L. MALOY (20)
CHIEF OF NAVAL ED. &
TRAINING
NAVAL AIR STATION
PENSACOLA, FL 32508

DR. WILLIAM E. NORDBROCK
FMC-ADCO BOX 25
APO, NY 09710

dr. jim hollan
code

DR. NORMAN J. KERR
CHIEF OF NAVAL TECH. TRAINING
NAVAL AIR STATION MEMPHIS (75)
MILLINGTON, TN 38054

DR. JAMES McBRIDE
NAVY PERSONNEL R&D CENTER
SAN DIEGO, CA 92152

Library, Code P201L
Navy Personnel R&D Center
San Diego, CA 92152

DR. JIM HOLLAN
CODE 14
NAVY PERSONNEL R&D CENTER
SAN DIEGO, CA 92152

DR. LEONARD KROEKER
NAVY PERSONNEL R&D CENTER
SAN DIEGO, CA 92152

DR. WILLIAM MONTAGUE
NPRDC CODE 13
SAN DIEGO, CA 92152

TECHNICAL DIRECTOR
NAVY PERSONNEL R&D CENTER
SAN DIEGO, CA 92152

H. WILLIAM GREENUP
EDUCATION ADVISOR (E031)
EDUCATION CENTER, MCDEC
QUANTICO, VA 22134

SPEC. ASS. FOR MARINE MATTERS
CODE 100M
OFFICE OF NAVAL RESEARCH
800 N. QUINCY STREET
ARLINGTON, VA 22217

TECHNICAL DIRECTOR
U.S. ARMY RESEARCH INST.
BEHAVORIAL & SOCIAL SCIENCES
5001 EISENHOWER AVENUE
ALEXANDRIA, VA 22333

DR. BEATRICE J. FARR
U.S. ARMY RESEARCH INST.
5001 EISENHOWER AVENUE
ALEXANDRIA, VA 22333

DR. HAROLD F. O'NEIL, JR.
DIRECTOR, TRAINING RES. LAB
ARMY RESEARCH INSTITUTE
5001 EISENHOWER AVENUE
ALEXANDRIA, VA 22333

MR. ROBERT ROSS
US ARMY RESEARCH INST. FOR THE
SOCIAL & BEHAVIORAL SCIENCES
5001 EISENHOWER AVENUE
ALEXANDRIA, VA 22333

DR. ROBERT J. SEIDEL
US ARMY RESEARCH INST.
5001 EISENHOWER AVINUE
300 N. WASHINGTON ST.
ALEXANDRIA, VA 22333

DIR., OFFICE OF MANPOWER
UTILIZATION
HQ. MARINE CORPS (MPU)
BCB, BLDG. 2009
QUANTICO, VA 22134

DR. A.L. SLAFKOSKY
SCIENTIFIC ADVISOR (CODE RD-1)
HQ, U.S. MARINE CORPS
WASHINGTON, DC 20380

MR. JAMES BAKER
ARMY RESEARCH INSTITUTE
5001 EISENHOWER AVENUE
ALEXANDRIA, VA 22333

DR. MYRON FISCHL
U.S. ARMY RESEARCH INST.
SOCIAL & BEHAVIORAL SCIENCES
5001 EISENHOWER AVENUE
ALEXANDRIA, VA 22333

COMMANDER, US ARMY RES. INST.
FOR BEHAVIORAL & SOCIAL SCI.
ATTN: PERI-BR(DR. J. ORASANU)
5001 EISENHOWER AVENUE
ALEXANDRIA, VA 22333

DR. ROBERT SASMOR
US ARMY RES. INST. FOR THE
BEHAVIORAL & SOCIAL SCIENCES
5001 EISENHOWER AVINUE
ALEXANDRIA, VA 22333

DR. HILDA WING
ARMY RESEARCH INST.
5001 EISENHOWER AVENUE
ALEXANDRIA, VA 22333

Headquarters, US Marine Corps
Code MPI-20
Washington, DC 20380

MAJOR FRANK YOHANNAN, USMC
HEADQUARTERS, MARINE CORPS
(CODE MPI-20)
WASHINGTON, DC 20380

DR. KENT EATON
ARMY RESEARCH INSTITUTE
5001 EISENHOWER BLVD.
ALEXANDRIA, VA 22333

DR. MILTON S. KATZ
TRAINING TECHNICAL AREA
U.S. ARMY RESEARCH INST.
5001 EISENHOWER AVENUE
ALEXANDRIA, VA 22333

JOSEPH PSOTKA, PH.D
ATTN: PERI-IC
ARMY RESEARCH INSTITUTE
5001 EISENHOWER AVENUE
ALEXANDRIA, VA 22333

DR. JOYCE SHIELDS
ARMY RESEARCH INST. FOR THE
BEHAVIORAL & SOCIAL SCIENCES
5001 EISENHOWER AVENUE
ALEXANDRIA, VA 22333

DR. ROBERT WISHER
ARMY RESEARCH INST.
5001 EISENHOWER AVE.
ALEXANDRIA, VA 22333